Cairo University Predictive Analytics

Faculty of Computers and Artificial Intelligence Spring 2023

Operations Research and Decision Support Department Lab #9

**Continue Regression Analysis**

**Lab Objectives**

* **Multiple Regression Model**
* They are extensions to the simple linear model and allow the creation of models with several independent variables

𝒀=𝜷𝟎+𝜷𝟏𝑿1+𝜷2𝑿2+…+𝜷k𝑿k

* + - 𝒀 = dependent variable (response) or Actual value
    - 𝑿i = ith independent variable (explanatory or predictable)
    - 𝜷𝟎 = intercept “value of Y when all Xi is zero”
    - 𝜷i = coefficient of the ith independent variable
    - k = number of independent variables
* Testing model significance
  + For each parameter (independent variable), to reject null hypothesis (the model/ parameter is significant): 𝑃−𝑉𝑎𝑙𝑢𝑒< ∝
* Adjusted r2 in Multiple Linear Regressions:
  + The adjusted r2 takes into account the number of independent variables in the model.
  + The adjusted r2 value is often used to determine the usefulness of an additional variable.

**Question 1:**

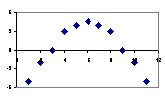
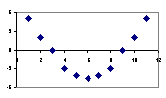
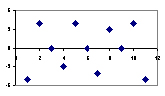
For a certain data set, a linear regression model was developed with 2 independent variables (with adjusted r2=0.6115). Then one more independent variable was added, that we are sure has no effect on the dependent variable. What do you think the adjusted new r2 value will be?

1. Larger than old r2
2. The same
3. Smaller than old r2

**Answer:**

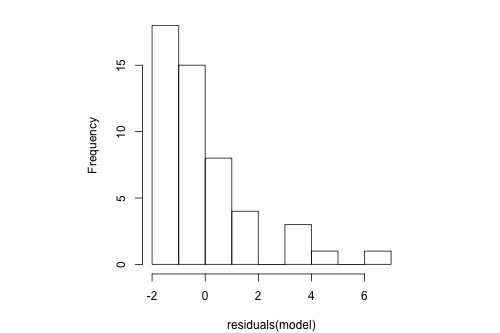
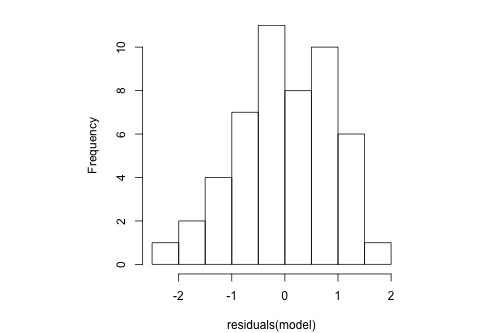
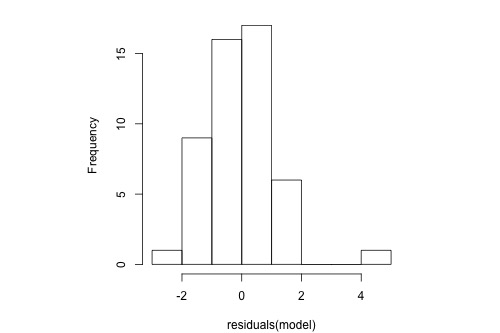
3, As the new r2 will only increase or stay the same if the newly added independent variable is significant to y

* **There are four assumptions associated with a linear regression model:**
  + Linearity:
    1. If the points in *a residual plot* are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a nonlinear model is more appropriate.



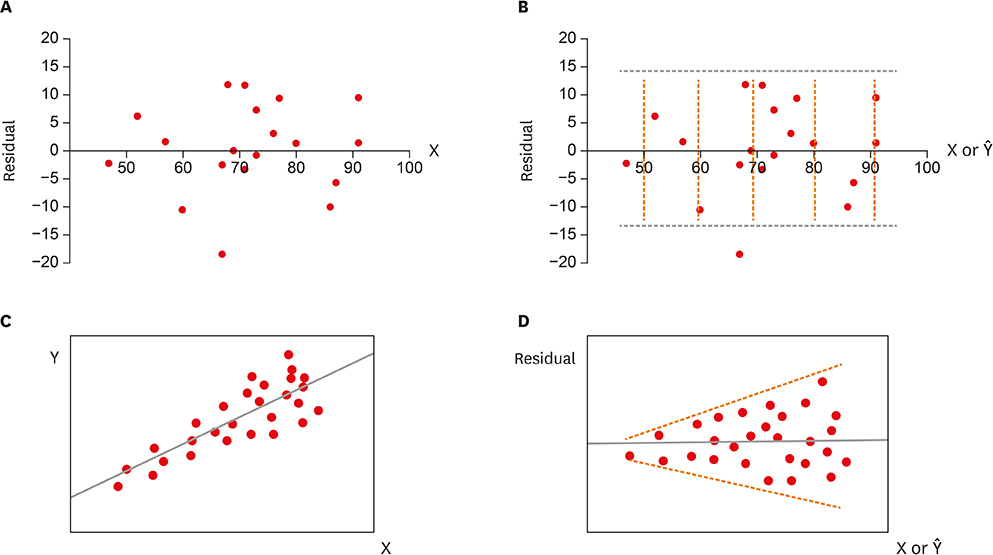
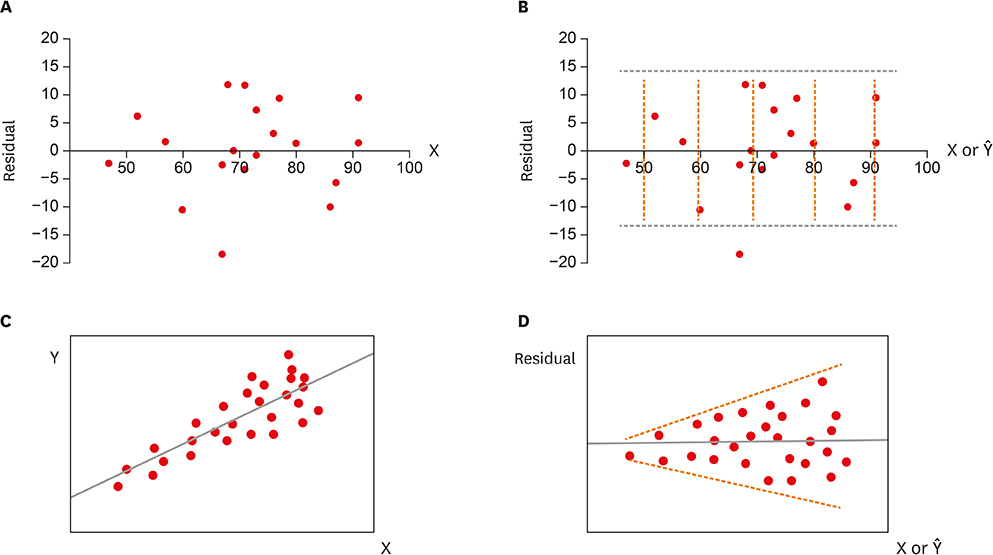
A random pattern, Non-random plot patterns,   
indicating a good fit for a linear model suggesting a better fit for a nonlinear model

* + Normality: The errors are normally distributed, with a mean of zero.
    1. If the *histogram* of standardized residuals is symmetric bell-shaped (evenly distributed around zero), then the normality assumption is likely to be true.
    2. If the *histogram* indicates that random error is not normally distributed, it suggests that the model's underlying assumptions may have been violated.



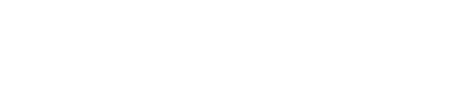
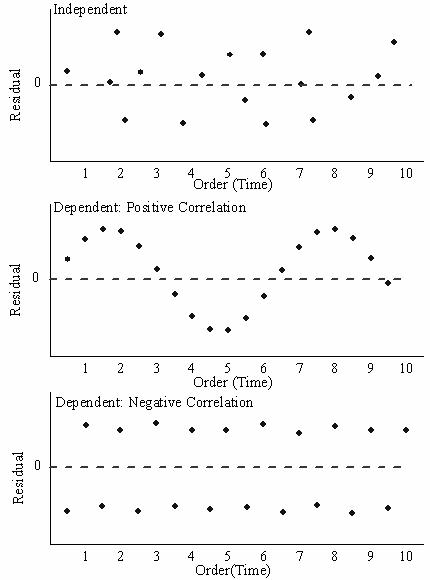
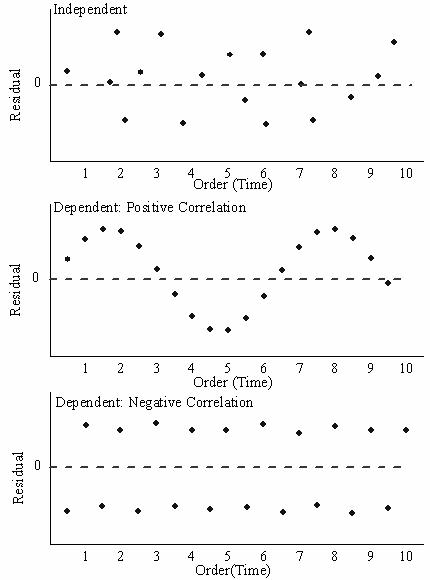
The Residuals are Residuals are normally distributed Residuals are normally distributed But, there is one extreme outlier not normally distributed (skewed)

* + Homoscedasticity: The residuals have a constant variance regardless of the value of X
    1. Data are homoscedastic if the *residuals plot* is the same width for all values of the independent variable.
    2. Heteroscedasticity is when the variability in the response is changing as the predicted value increases.



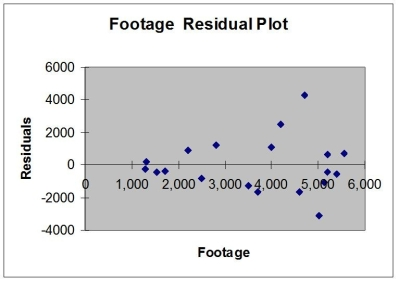
Residual plot showing homoscedasticity Residual plot showing heteroscedasticity

* + Independence: the distribution of errors is random and not influenced by or correlated to the errors in prior observations.



**Question 2**

Based on the residual plot below, you will conclude that there might be a violation of which of the following assumptions?

****

A) independence of errors C) linearity of the relationship

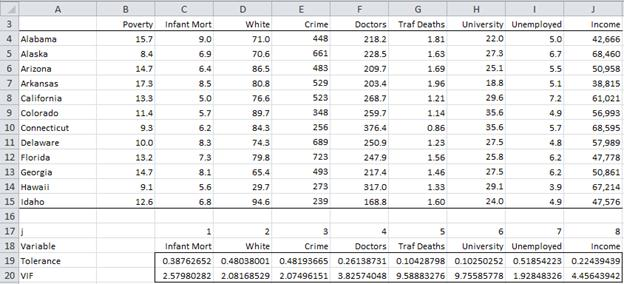
B) homoscedasticity D) normality of errors

**Ans. B**

* **Multi-collinearity**
* Multi-collinearity occurs when two or more explanatory variables are highly correlated to each other.
  + This mean that they do not provide unique or independent information in the regression model (redundancy).
  + If the degree of correlation is high enough between explanatory variables, it can cause problems when fitting and interpreting the regression model.
* It’s possible to detect multi-collinearity using a metric known as the **variance inflation factor (VIF)**, which measures the correlation and strength of correlation between the explanatory variables in a regression model.
  + It is basically calculated by performing a multiple linear regression using one explanatory variable as the response variable, and the other(s) as the explanatory variables.
  + This variable’s VIF is calculated as 1 / (1 – R Square)
  + Then this process is repeated for the other variable(s)
* The value for VIF starts at 1 and has no upper limit. A general rule of thumb for interpreting VIFs is as follows:
  + A value of 1 indicates there is no correlation between a given explanatory variable and any other explanatory variables in the model.
  + A value between 1 and 5 indicates moderate correlation between a given explanatory variable and other explanatory variables in the model, but this is often not severe enough to require attention.
  + A value greater than 5 indicates potentially severe correlation between a given explanatory variable and other explanatory variables in the model. In this case, the coefficient estimates and p-values in the regression output are likely unreliable.

**Question 3**

The US Census Bureau collects statistics comparing the various 50 states. The data includes the poverty rate (% of population below the poverty level) and infant mortality rate per 1,000 live births) by state, in addition to a number of other statistics. Data for the first 12 states are displayed below. Which variables should we be concerned about?



**Ans. We should be concerned about the Traffic Deaths and University variables as their VIFs are greater than 5.**

* **Confidence Interval for a Regression Coefficient**
* We are usually working with sample data, a few data points, instead of the whole population of interest.
* For example, if you want to get the average height of teenage girls in a middle school, you get data for a few girls and use it to estimate the average of all the girls.
* Similarly, the line of best fit (𝒀=𝜷𝟎+𝜷𝟏𝑿1+𝜷2𝑿2+…+𝜷k𝑿k) is computed from a random sample of measurements of Xs. If we used a different dataset, we would most likely compute slightly different values for the 𝜷 parameter. Thus our values are always estimates and as such have a confidence interval associated with them.
  + Confidence intervals in regression problems are calculated with the formula

()

* + - b is the predicted value of the regression coefficient,
    - is the value from the t table with confidence level, and degrees of freedom, df= n-k-1,
    - is the standard error for b

**Question 4:**

Using the data in the lab’s Excel file, we can analyze a given state’s gross production.

We would like to run a regression with the Gross State Product as the dependent variable, and Capital Input and Labor Input as independent variables:

1. Perform basic data diagnostics (linearity, normality, homoscedasticity, multi-collinearity)
2. Check for the significance of the model and its goodness-of-fit. Interpret your results.
3. Using the resulted coefficients, write the regression equation and comment on your findings.
4. Let βj denote the population coefficient of the jth regressor (intercept, Net Capital Stock, and Hours Worked). Get the 95% confidence interval for slope coefficient β2 from Excel output.

**Answer:**

1-

* **Linearity**
  + **Residual Plots: the points in the residual plots are randomly dispersed around the horizontal axis**
* **Normality: the residual distribution histogram shows that residuals follow normal distribution pattern.**
* **Homoscedasticity: the data are homoscedastic, as none of the residual scatterplots shows a clear cone-shaped pattern.**
* **Multi-collinearity: VIF & Correlation matrix indicate a multi-collinearity problem**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Multi-collinearity Checking** | | | | |
|  | **VIF** |  | | | |
|  |  |  | | | |
|  | 13.18395084 |  | | | |
|  |  |  | | | |
| ***Correlation Matrix*** | | | **Gross state product ($ Millions) Y** | **Net capital stock ($ Millions) x1** | **Hours worked in all jobs ('000 Hours) x2** | |
| **Gross state product ($ Millions) Y** | | | 1.000 | 0.980 | 0.991 | |
| **Net capital stock ($ Millions) x1** | | | 0.980 | 1.000 | 0.961 | |
| **Hours worked in all jobs ('000 Hours) x2** | | | 0.991 | 0.961 | 1.000 | |

2-

* **The standard error of the regression and R-squared are two key goodness-of-fit measures for regression analysis. Note that k=2, as there are two explanatory variables used in the regression.**
* **The multiple correlation coefficient 0.9960 indicates that the correlation among the independent and dependent variables is positive**
* **R square = 0.9920 means that close to 100% of the variation in the dependent variable (Gross State Product) is explained by the independent variables (Net Capital Stock and Hours Worked)**
* **Since the p value (1.07142E-21) is smaller than α=0.05, we reject the null hypothesis that the variables are unrelated. In other words, there is a relation between the dependent variable and independent variables**

3-

* **The regression equation can be written as: Y = -92349 + 0.084 x1 + 0.072 x2**
* **A 10 percent increase in Net Capital Stock, should lead to a 0.8% increase in Gross State Product**
* **A 10 percent increase in Hours Worked, should lead to a 0.7% increase in Gross State Product**

4-

* **The 95% confidence interval for slope coefficient β2 from ANOVA table is: (0.055,0.089) ->0.07**